1

Experiments as Fair Tests

Causal claims abound in everyday life. Open your medicine cabinet: Do these over-the-counter products to ameliorate headaches or bug bites actually work as advertised? Open your closet: Does your choice of wardrobe change the way that others treat you? Open your refrigerator: Did you buy nonfat milk because you hoped to lose some weight? Open today's news: Do government programs have a material effect on economic or social outcomes? Stepping back to reflect for a moment, you probably have hunches about whether these pills or products or policies change the way that things turn out.

How did you come to acquire these beliefs about cause and effect? In some cases, you may have drawn on your own experience and intuitions. Did your allergy symptoms subside when you took that hay fever medication? Without putting too much thought or effort into settling the question, a reasonable person might conclude that the medication worked because it produced an abrupt change in symptoms. Where you lacked direct personal experience, you may have drawn inferences based on other things that you know. For example, you might infer that low-fat milk will help you lose weight because it contains fewer calories than whole milk. In addition to drawing your own conclusions based on experience and logic, you may have accepted other people's causal assessments, such as recommendations by health professionals or policy evaluations by knowledgeable researchers. Your causal beliefs may reflect a combination of all three sources.

Informal empiricism sometimes works quite well. Experience is probably an adequate guide if your aim is to figure out whether staying up all night makes you drowsy the next day or whether turning on the air conditioner reduces the temperature in your room. Reasoning based on background knowledge may be a reliable guide on questions such as whether it is safe to operate a corded electrical appliance outdoors in the rain. And experts may provide reliable guidance on questions such as the share of the population that must be vaccinated in order for "herd immunity" to prevent the spread of infectious disease.

On the other hand, informal empiricism may give us false confidence in causal claims. Many people swear that eating local honey relieves seasonal allergies. In addition to offering testimonials from those whose allergy symptoms subsided after eating local honey, proponents of this hypothesis argue that this effect can be deduced from the fact that eating local pollen, which is mixed into the honey, helps allergy-sufferers build up immunities. However, skeptics question whether honey-borne pollen can be expected to

help alleviate allergy symptoms. Local honey may contain the wrong kind of pollen – as one summary points out, "Bees typically pollinate colorful flowers, yet most people are allergic to pollen released by grasses, trees and weeds."[1] Informal empiricism may lead people to draw opposing conclusions about cause and effect.

We often look to experts to resolve such controversies, but experts, too, are fallible. They are often overconfident (Dawes 2009), and their pronouncements may be tinged by conflicts of interest (Tetlock 2017). For example, political campaigns in the United States spend vast sums of money on digital advertising, and consultants who sell these ads are quick to vouch for their effectiveness, pointing to campaigns that skillfully used digital ads on their way to victory. But carefully controlled experiments cast doubt on whether these ads persuade their target audience to change their vote (Broockman and Green 2014). In clinical medicine, physicians have been found to continue to perform surgical procedures even after they have been shown to be ineffective. Arthroscopy is commonly used to treat knee arthritis, despite the fact that a carefully-controlled experiment showed the procedure to be no more effective than a placebo operation (Gerber and Patashnik 2006; Siemieniuk et al. 2017). Experts also proved unable to predict which of 54 messages to encourage exercise would get gym members to visit their gym more often. A randomized trial involving more than 60,000 gym members showed that only a small number of encouragement messages produced positive results and that experts grossly overestimated the effectiveness of every message (Milkman et al. 2021).

The preceding examples underscore the importance of research design when adjudicating questions of cause and effect. Take the example of SAT preparatory courses. Full-blown courses can be quite expensive, and each year the test-prep industry takes in billions of dollars. What evidence suggests that in-person classes or online instruction raises SAT scores substantially? First, test-prep companies point out that those who take the courses typically receive much higher scores than those who do not. The superior performance of course-takers bolsters the intuition that hard work preparing for the test pays off. But skeptics question this interpretation, pointing out that high-achieving high school students are more likely to enroll in a prep course than those with less distinguished academic records. Even if the courses truly had no effect on SAT scores, we would still expect higher scores among those who enroll in these courses than those who do not. A second body of evidence shows that those who spend more hours studying for the SAT receive higher scores than their counterparts who spend fewer hours. Again, skeptics object that high-achieving students tend to invest long hours in test preparation, so, even if studying had no effect on scores, we would expect higher scores among those who studied extensively. A third body of evidence suggests that SAT scores are higher among people who take prep courses than among those who do not, even when the comparison focuses solely on high school students with similar grades and from similar schools. This type of detailed comparison is more persuasive than a coarse comparison

[1] Quoted in www.aentassociates.com/raw-local-honey-cure-allergies/ (last accessed June 8, 2021). The scientific evidence is at best mixed. One study conducted in Connecticut (Rajan et al. 2002) found local honey to have no effects on allergy symptoms; another in Malaysia (Asha'ari et al. 2013) found equivocal evidence. I thank Dr. David Gudis and Dr. Stephen Canfil for their insights into the theory and evidence concerning this hypothesis.

of those who do or do not take prep classes, but it still leaves room for skepticism. Why did certain high school students take the prep class? Two students from the same school can have the same grades but different levels of ability or motivation. If the more able or motivated are more likely to take the prep course than their counterparts, the course may appear to work even if it truly has no effect. A fourth body of evidence focuses on the fact that students who enroll in a prep course after receiving a disappointing initial SAT score tend to do better on their second try at the test. Test prep companies point to this pattern as proof that their courses boost test scores. Skeptics, however, point out that this pattern merely reflects the fact that people who receive disappointing scores will, on average, score higher next time (just as people who are surprised to receive strong scores will tend to do worse if they retake the exam). Finally, what about evidence showing that, among people who received the same initial SAT score, those who took a prep course before taking the test again tended to score higher on their second try than those who did not take a course? Same problem, says the skeptic. Imagine two people with the same initial score. Disappointed and expecting to improve, one student takes a prep course; satisfied and perhaps a bit complacent, the other student does not. Even if the course truly has no effect on the second test, one might still expect the first student to receive a higher score than the second student.

Notice that the back-and-forth between prep course proponents and skeptics has a recurrent theme. Any evidence that proponents offer will be rejected by skeptics on the grounds that those who take preparatory courses have higher scholastic aptitude to begin with than those who do not take the courses.

What kind of evidence would satisfy a determined skeptic? Presumably, the skeptic wants to see what happens when students' scholastic aptitude is unrelated to whether they take a prep course or not. In that case, the skeptic cannot complain that those who took the course were more able or motivated to begin with. But where would we find such a comparison?

One approach is to compare students who decided to take the class or not, focusing just on the students who seem to share identical background attributes. They appear to have the same grades; they come from the same schools; their parents have the same educational credentials; and so on. On the surface, those who took the prep classes look just like those who did not. But the skeptic still wonders whether these two groups are truly similar. Yes, they may have the same observed attributes, but what about the attributes that were not measured, such as their level of motivation? Very well, suppose researchers were to measure students' motivation, perhaps by administering a survey, and focus the comparison even more narrowly on those who seem to have the same level of motivation. Nevertheless, other attributes remain unmeasured, such as the encouragement they receive from their peers. Even when presented with "big data" – a study involving vast numbers of students and copious information about each student's background – the skeptic will continue to wonder whether there remains some unmeasured attribute that leads certain students to enroll in prep courses and perform especially well on the SAT.

An alternative approach, and the focus of this book, is to randomly assign students to either take an SAT prep class or not. For example, many high school students want to take a prep course but come from families that cannot afford to pay for it. Imagine that

these eager students enroll in a lottery and winners receive tuition-free admission to prep classes immediately. Both groups receive free admission to an SAT test to be administered three months from now. Suppose that all the lottery applicants (both winners and losers) take this test. And suppose that the lottery winners indeed attend their prep course, while the lottery losers prepare as they ordinarily would, without the benefit of a course. Would this research design satisfy a determined skeptic? The skeptic's earlier objection stemmed from the concern that students ordinarily self-select into taking a prep course, but in this research design, known as a *randomized experiment*, luck of the draw differentiates lottery winners from the lottery losers. Assuming the lottery is conducted honestly, there is no reason to expect that more gifted or determined students will win admission to the test-prep course. By chance, one group might be more determined or gifted than the other before the class begins, but there is no reason to expect the lottery winners to have more scholastic aptitude than the lottery losers.

Experiments are characterized as fair tests because, when the experimental groups are formed by chance, neither group has a head start over the other. In this example, if the group that attended the prep course received higher average scores than the group that did not attend the prep course, the skeptic must come to grips with this stubborn fact. Of course, the skeptic may still harbor doubts. No experiment is perfect. Were the lottery losers demoralized because they were unable to enroll in a prep course? Perhaps higher scores among the lottery winners reflect not the benefits of the course per se but demoralization among the lottery losers? Or perhaps by chance more able students were selected as lottery winners? Conducting a fair test does not necessarily end debate, but now the scope of debate narrows considerably. Anticipating the criticism that lottery losers become demoralized, a researcher could conduct a survey of all lottery participants shortly before the exam. If morale appears to be the same among lottery winners and losers, this criticism seems unfounded. If the concern is that random chance assigned especially able students to the test-prep group, this conjecture can be evaluated statistically using the methods described in the chapters that follow. Or researchers can repeat the experiment until the it-happened-by-chance conjecture becomes untenable. Eventually, fair tests will home in on the answer to the question of whether SAT prep classes work. As experimental evidence mounts, it will also shed light on the question of how well they work, for whom, and under what conditions.

Experiments are especially important when the "treatment" – a pill, a prep course, an experience – is something that people ordinarily select for themselves. Consider the experiment conducted by Balcells et al. (2022) on the effects of bringing Chilean university students to the Museum of Memory and Human Rights, which memorializes victims of General Augusto Pinochet's dictatorship in Chile. Students who had not previously visited the museum were invited to meet on campus and were randomly assigned to visit the museum or simply complete an exit survey. Before attending the museum, the 126 students who later attended the museum expressed the same average level of support for military rule as the 106 students in the control group. When all participants were reinterviewed a week later, 17 percent of students who attended the museum supported military rule, as compared to 28 percent of their control group counterparts, a difference of 11 percentage points.

What if this study had not randomly assigned students to visit the museum, but instead simply surveyed those who attended the museum? First, absent a control group, it would not be clear what their level of support for military dictatorship should be compared to. Second, absent random assignment to treatment and control, it would be unclear whether the museum experience was a cause of attitude change. For example, a survey that merely compared the opinions of people who attended the Museum of Memory and Human Rights to the opinions of people who attended a nearby art museum would leave us wondering whether people with different views about military rule chose to go to different museums.

1.1 AN OVERVIEW OF THIS BOOK

The chapters that lie ahead are written with several goals in mind. The first is to urge you to think critically about the causal claims that you encounter in everyday life. What evidence supports or refutes a causal claim? The second goal is to convey a sense of why experiments often provide convincing evidence about cause-and-effect. Why are experiments so often characterized as the "gold standard" for evaluating whether an intervention is effective? The third goal is to prepare you to read experiments that illuminate important cause-and-effect relationships in psychology, education, politics, economics, health, crime, media studies, and many other fields. The fourth goal is to appreciate the limitations of an experimental design or the way an experiment was implemented. What changes would have made the study more convincing or useful? The final goal is to inspire you to look at the world from an experimental vantage point. When the time comes to make an informed decision, you may need to read or conduct an experiment, and this book helps you think about the ingredients that make for an informative experimental design.

To make the material come to life, this book takes a hands-on approach. To consolidate your understanding of the material, you will be conducting your own experiments. Chapter 2 lays out the key ideas that inform an experimental design, and Chapter 3 gives you an opportunity to design, implement, and analyze a small product-testing experiment. Chapter 4 provides a panorama of social science experiments conducted in a variety of countries on topics ranging from reducing prejudice to preventing deforestation. Chapter 5 lays out ethical issues to consider before conducting an experiment involving human participants. Chapter 6 walks you through the process of planning a social science experiment, implementing the experimental design, and addressing contingencies that may arise during data collection. This chapter also guides you through the process of writing up your statistical analysis and archiving your experimental materials when your study is complete. Chapter 7 invites you to learn some graphical and statistical methods that will enable you to do more with your data. If you are reading this book as part of a statistics course, experiments will help deepen your understanding of statistics, and vice versa.

2

Key Terms

This chapter introduces key terms used to describe experiments and, more generally, the investigation of cause and effect. Because so many different disciplines use experiments, layers of overlapping terminology have accumulated, and this chapter tries to cut through the clutter by grouping synonyms, thereby keeping jargon to a minimum. In addition to providing definitions, this chapter explains why these key concepts are important in practice.

This chapter aspires to be more than a glossary of terms. It presents definitions in a logical sequence that starts with the basic ingredients of an experiment (treatments, outcomes), takes up the question of what we mean by a causal effect, and then builds to the core assumptions on which experiments depend.

2.1 TREATMENT OR INTERVENTION

In biomedical research, the term *treatment* refers to a medical procedure or pharmaceutical product whose effects are to be evaluated. For example, one might want to know whether a specific type of stomach reduction surgery known as gastric bypass leads to long-term weight loss. Briefly put, this surgery reduces the size of the stomach so that patients will feel full sooner when eating. Medical professionals have described this treatment in great length, even going so far as to film the procedure.[1]

When describing any treatment, it is important to clarify what counts as the absence of treatment. For example, will gastric bypass surgery be contrasted with another surgical procedure designed to promote weight loss? Or contrasted with a sham surgery that does not affect the digestive tract? Or contrasted with no surgery at all? This clarification about the *control* condition can have important implications for the design of the study and the interpretation of the results. If we contrast gastric bypass surgery with sham surgery (assuming the patient is unaware of which kind of surgery actually took place), we are testing the physiological effects of this specific procedure while holding constant the psychological effects of believing that one has undergone a

[1] See https://asmbs.org/patients/bariatric-surgery-procedures (last accessed June 10, 2021).

stomach reduction operation. Studies of this kind are called *placebo–control trials* because the inactive placebo leads subjects to believe that they are receiving the treatment. The administration of a placebo also allows the researcher to compare only people who demonstrate a willingness to undergo surgery; some of these subjects receive gastric bypass surgery, and others receive sham surgery.

On the other hand, if we contrast gastric bypass surgery to no surgery whatsoever, we reveal the combined physiological and psychological effects of the procedure, assuming everyone who is slated to receive a gastric bypass actually goes through with the surgical procedure. Usually, biomedical researchers are primarily interested in the physiological effects of a surgical procedure, not the psychological effects of undergoing any surgery, such as a sham surgery in which the patient is sedated, an incision is made, and the wound is stitched back together.[2]

The term "treatment" is used outside the domain of biomedical research to refer to a wide array of interventions whose effects are the object of study. Here are some illustrative examples of interventions that have been studied by social scientists:

- Economists study the effects of social programs that provide cash assistance to the poor (Baird et al. 2014). The intervention is the new program of cash assistance, and the control group receives only the usual support from social welfare programs.
- Criminologists study the effects of increased police surveillance in high-crime neighborhoods (Collazos et al. 2021). The intervention is a large increase in the amount of time that police patrols spend in designated locations; the control locations receive the usual level of police presence.
- Psychologists study the effects of an online diversity training program for corporate employees (Chang et al. 2019). The intervention group is exposed to an hourlong online training; the control group does not receive this training (but may be exposed to the training sometime in the future).
- Political scientists study the effects of mailings designed to encourage people to vote in an election (Panagopoulos 2011). The intervention in this case is a letter that urges registered voters who have voted in the past to cast a ballot in the upcoming election and thanks them for voting in a recent election. The control group receives no letter.[3]

Notice that in some cases the interventions are very specific – a particular diversity training program or a particular get-out-the-vote mailer. In other cases, the intervention is characterized in general terms. Increased police surveillance in the Collazos et al. (2021) study meant that, on average, police spent 50 percent more time in treated locations than they did in untreated locations, but the actual increment of police attention varied from one treated location to the next. In other words, interventions may be defined narrowly and implemented in a homogeneous manner, or they may be

[2] Their preference for placebo-control designs may also reflect a practical consideration. If patients slated to receive surgery frequently decide not to go through with it, the no-surgery-control design introduces an additional layer of complexity because it is unclear which subjects in the control group would have gone through with the surgery had they been slated to receive it. This kind of "failure-to-treat" problem falls outside the scope of this book but is discussed at length in chapter 5 of Gerber and Green (2012).

[3] This study also had another control group that received a letter reminding voters of the upcoming election but made no reference to past participation. See Exercise 2.3.

defined more loosely so that the treatment refers to a set of interventions that share the same basic characteristics (for example, they all involve cash transfers).

2.2 SUBJECTS OR PARTICIPANTS

Every social experiment involves some set of observations, whether they be people, firms, schools, media markets, or countries. When the observations are people, they are traditionally called *subjects*, although many researchers nowadays refer to them as *participants*.[4] In the get-out-the-vote study, the subjects are registered voters. Researchers send some subjects a "thank you for voting" mailing and assess whether this intervention makes them more likely to vote than the control group that receives no mail.

In other experiments, determining who the subjects are may be a bit tricky. In the policing study, one might say that the subjects are block-long stretches of road called street segments. Some segments receive extra police attention, while others receive the usual amount, and the question is whether additional police surveillance causes the treatment segments to experience less crime than the control segments. However, when we change the focus to how residents of treated and untreated segments feel about crime, one might think of the residents as subjects; some residents have been treated with extra police presence and others not. This example calls attention to a subtlety: In order to discern who or what the subjects are, we must first specify what the experimental outcomes are. A study with a variety of outcomes might have more than one set of subjects.

2.3 OUTCOMES

When researchers conduct experiments, they often do so with a concrete outcome in mind. For example, to what extent do get-out-the-vote mailings increase the rate at which recipients vote? In this case, the mailings constitute the treatment, and voting constitutes the outcome. Every participant in this study is associated with an outcome, either "voted" or "did not vote."

Sometimes the task of measuring outcomes requires an extra step: moving from an idealized outcome to an outcome that can be measured in practice. For example, researchers might wonder how cash transfers to poor families affect health outcomes. Are children whose families receive cash subsidies healthier than those whose families receive no special assistance? The outcome is health, but what does it mean to be healthy? One can imagine a variety of definitions, such as physical vigor or the absence of disease, and scholars actively debate the merits of competing definitions.

But even if researchers could agree on a single definition, there remains the challenge of measuring health. For example, if health were defined as the absence of disease, how extensively would researchers have to search for disease before declaring it to be absent? The constraints of time and money often dictate what researchers can measure in practice. Some studies of cash transfers simply measure the height and weight of each child as a *proxy* for health. Here, the term proxy refers to a stand-in for the underlying outcome variable that

[4] The term "subjects" may also refer to nonhuman participants, such as plants or products. Subjects are also referred to as *units* or *observations*.

Box 2.1 Methods to Assess Whether a Proxy Adequately Measures an Underlying Concept

When social scientists assess the validity of an outcome measure, they are evaluating whether a proposed measure gauges a particular underlying concept and nothing more. For example, the validity of the body mass index (BMI) as a measure of health boils down to whether an increase in health implies an increase in BMI and whether a decrease in health implies a decrease in BMI.

Social scientists typically use three methods for validating outcome measures:

The first is to assess the *face validity* of the measure. Does the measure bear an intuitive relationship to the underlying concept? In the case of BMI, it makes intuitive sense that a person who contracts a debilitating illness that prevents them from eating will experience declining body mass. On the other hand, one can also think of intuitive scenarios in which a person becomes morbidly obese, in which case they are less healthy but higher in BMI. The face validity of BMI as a health measure is therefore ambiguous.

Second, *criterion validity* considers whether the proposed measure confirms the judgements of experts or other carefully calibrated measures. For example, suppose one were to gather a set of people whom physicians declare to be very ill and another set whom physicians declare to be in excellent health. Would the first group have low BMI scores and the second group have high BMI scores? One reason why BMI is thought to be a valid measure of health among children in low-income countries is that, in these locations, BMI scores tend to coincide with physicians' assessments based on extensive medical examinations.

Third, *discriminant validity* asks whether the proposed measure is distinct from concepts that are not the intended target of measurement. For example, in a location where different ethnic groups tend to differ in height, BMI might reflect ethnicity rather than health. This evidence might call into question the validity of the BMI measure, or it might imply that BMI is valid when comparing members from a given ethnic group but is not valid when comparing people from different ethnic groups.

Finally, even when measures are valid, they may nonetheless contain some degree of measurement error. For example, a tape measure is a valid indicator of length (why?), but in practice those who use tape measures make errors when reading or transcribing numbers.

the researcher seeks to measure. When reading experimental research, it is important to attend to the distinction between the concepts that researchers strive to measure and the proxies that they use as actual measures. Whether a proxy convincingly measures the underlying concept of interest is a matter of interpretation. See Box 2.1 for a summary of methods to assess whether a proxy adequately measures an underlying concept.

When assembling the data from an actual study, a researcher records the outcome for each subject. For example, a study on cash subsidies and health may record the body mass index (BMI) for each child.[5] For concreteness, consider the outcomes for five children from different households. Two children are from families that received cash and three from families that did not, as shown in Table 2.1. As you inspect this table,

[5] In the metric system, BMI is defined as weight in kilograms divided by height in meters squared. To calculate BMI using imperial measures, the formula is `bmi <-(weight_pounds/height_inches^2)*703`.

TABLE 2.1. *Hypothetical schedule of potential outcomes for five subjects*

Subject no.	(1) Observed treatment	(2) Observed outcome	(3) Treated potential outcome*: Y(1)	(4) Untreated potential outcome*: Y(0)	(5) Treatment effect**	(6) Covariate: BMI measured 1 year earlier
1	1	16	[16]	11	5	13
2	1	20	[20]	12	8	12
3	0	25	25	[25]	0	24
4	0	29	27	[29]	−2	28
5	0	18	22	[18]	4	23
			Average Y(1)	Average Y(0)	ATE	Average covariate
			22	19	3	20
			Boxed entries observed*	Boxed entries observed*	No entries observed**	
		Average outcome among those treated: 18				
		Average outcome among those not treated: 24				
		Difference-in-means estimate: −6				

notice that the BMI outcomes tend to be lower for children in families that received cash subsidies. Did the cash subsidy somehow make children smaller? Or did cash tend to go to families whose children were smaller to begin with? In order to distinguish the effect of cash subsidies on BMI from the apparent relationship between cash subsidies and BMI, we must first clarify what we mean by a treatment effect. To do so, we introduce the concept of potential outcomes.

2.4 POTENTIAL OUTCOMES

When we observe an outcome for a participant who received the treatment, we might wonder what would we have observed if this participant had not received the treatment? Conversely, when we observe an outcome for an untreated participant, we might wonder what would their outcome have been had they been treated? Unfortunately, we cannot observe participants in both their treated and untreated states; we can only observe one or the other. Still, we can imagine that every participant has two potential outcomes, one that would be expressed if they were treated and another that would be expressed if they were not treated. The *causal effect of a treatment for a given subject is the difference between these two potential outcomes*.

In order to get a clearer sense of the mapping between potential outcomes and observed outcomes, return to Table 2.1. Column (3) indicates the potential BMI outcomes that subjects would express if they were treated, and column (4) indicates the potential BMI outcomes that subjects would express if they were untreated. *In practice, researchers never get to see both of these columns; the numbers in Table 2.1 are offered as a hypothetical example to clarify some important concepts.* The five entries in boxes